هوش مصنوعی — تمرین نوشتاری شماره ۲

مدرس: **دکتر فدایی** — مهلت تحویل: پنج‌شنبه ۲۸ دی ۱۴۰۲، ساعت ۲۳:۵۹

احتمال

طراح: هادی بابالو

در یک مسابقه‌ی فوتبال ۴ تیم در مرحله‌ی گروهی حضور دارند. همه‌ی تیم ها دو به دو با هم بازی می‌کنند و هر دو تیم یک بازی با هم خواهند داشت. اگر بازی مساوی شود، دو تیم ۱ امتیاز خواهند گرفت و در غیر این صورت برنده ۳ امتیاز می گیرد و بازنده امتیازی نمی گیرد.

می‌دانیم احتمال برد تیم اول، برد تیم دوم و مساوی با هم برابر و مساوی $\frac{1}{3}$ است. تیم‌های a, b, c, d در یک گروه هستند و می‌دانیم در پایان مرحله گروهی تیم a توانسته ۶ امتیاز کسب کند. احتمال اینکه تیم b مرحله گروهی را با کسب ۴ امتیاز به اتمام رسانده باشد چقدر است؟

پاسخ)

پیشامد A را اینکه تیم a در نهایت 6 امتیاز کسب کرده باشد در نظر می‌گیریم.

همچنین پیشامد B را هم اینکه تیم b در نهایت 4 امتیاز کسب کرده باشد در نظر بگیرید.

در این صورت پاسخ مسئله برابر $P(B|A)$ است. از طرفی می‌دانیم که:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

پیشامد A فقط در صورتی رخ می‌دهد که تیم a دو بازی را ببر و یک بازی را ببازد.

$$P(A) = 3 \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

حال برای به دست آوردن $P(A \cap B)$ روی نتیجه بازی دو تیم a و b حالت‌بندی می‌کنیم. طبق نتایج حاصل شده در نهایت، مشخص است که نتیجه این بازی نمی‌توانسته تساوی باشد. پس کافی است صرفا دو حالت را بررسی کنیم.

$$P(A \cap B) = P(A \cap B \mid W_a) \times P(W_a) + P(A \cap B \mid W_b) \times P(W_b)$$

$$= \frac{1}{3} \times 2 \times \left(\frac{1}{3}\right)^2 \times 2 \times \left(\frac{1}{3}\right)^2 + \frac{1}{3} \times \left(\frac{1}{3}\right)^2 \times 2 \times \left(\frac{1}{3}\right)^2 = \frac{6}{243}$$

در نتیجه احتمالی که به دنبال آن هستیم برابر است با:

$$P(B \mid A) = \frac{\frac{6}{243}}{\frac{1}{9}} = \frac{2}{9}$$

شبکه‌های بیزی

طراح: هادی بابالو

می‌خواهیم عملکرد دانشجویان یک کلاس را در امتحان درس هوش مصنوعی پیشبینی کنیم. می‌دانیم دانشجویانی عملکرد خوبی در امتحان دارند که به خوبی برای آن مطالعه کرده باشند و همینطور در زمان امتحان سردرد نداشته باشند. از طرفی در نظر می‌گیریم که سردرد یا ناشی از خستگی و یا ناشی از سینوزیت است. فرض کنید مطالعه، سینوزیت و خستگی دو به دو از هم **مستقل** هستند.

در نظر بگیرید که متغیر F نشان‌دهنده سینوزیت داشتن، T نشانه خسته بودن، H نشانه سردرد داشتن، S نشان‌دهنده مطالعه و در نهایت E نشان‌دهنده امتحان را خوب دادن باشد. توجه کنید که تمام متغیرهای تعریف شده **باینری** هستند.

الف) شبکه بیزی متناظر با این مسئله را رسم کنید.

ب) توزیع شبکه بیزی رسم شده را بنویسید.

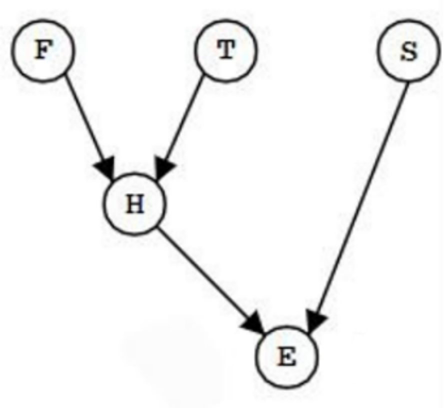
ج) احتمال اینکه سنا سینوزیت داشته باشد، خسته نباشد، سردرد داشته باشد و امتحانش را خوب ندهد را بر اساس احتمال‌های شرطی و به ساده‌ترین شکل ممکن بنویسید.

د) فرض کنید که در این بخش می‌دانیم که خستگی روی مطالعه تاثیر می‌گذارد. با این فرض و در شبکه بیزی جدید درستی یا نادرستی موارد زیر را با ذکر دلیل **مختصر** مشخص کنید.

- متغیرهای F و T به شرط دانستن H از هم مستقل هستند.
- متغیرهای E و F به شرط دانستن H از هم مستقل هستند.
- متغیرهای F و T به طور کلی از هم مستقل هستند.
- متغیرهای S و H به طور کلی از هم مستقل هستند.

پاسخ)

الف)



ب)

$$P(F) \times P(T) \times P(S) \times P(H \mid F, T) \times P(E \mid H, S)$$

ج)

$$P_{(+f,-t,+h,-e)} = \sum_{s} P_{(+f)} P_{(-t)} P_{(s)} P_{(+h|+f,-t)} P_{(-e|+h,s)}$$
$$= P_{(+f)} P_{(-t)} P_{(+h|+f,-t)} \sum_{s} P_{(s)} P_{(-e|+h,s)}$$

د)

- نادرست. پدران در حالت کلی نسبت به هم مستقل‌اند ولی به شرط دانستن فرزندشان از هم مستقل نیستند.
- درست. فرزندان از اجدادشان به شرط دانستن پدران مستقل هستند.
- درست. پدران در حالت کلی نسبت به هم مستقل‌اند
- نادرست: فرزندان به شرط دانستن پدران از هم مستقل هستند.

Hidden Markov Models

طراح: کیانوش عرشی، محمدطاها فخاریان

یک HMM با حالت‌های $\{a, b, c\}$ و الفبای $\{x, y, z\}$ تعریف کنید. احتمالات پایدار اولیه برای هر کدام از حالت‌ها برابر $\pi_a = 1$، $\pi_b = 0$ و $\pi_c = 0$ می‌باشند. همچنین احتمالات transition و emission در جدول زیر تعریف شده‌اند.

	a	b	c		x	y	z
a	0.2	0.8	0.0		0.8	0.2	0.0
b	0.0	0.8	0.2		0.0	0.6	0.4
c	0.4	0.0	0.6		0.2	0.0	0.8

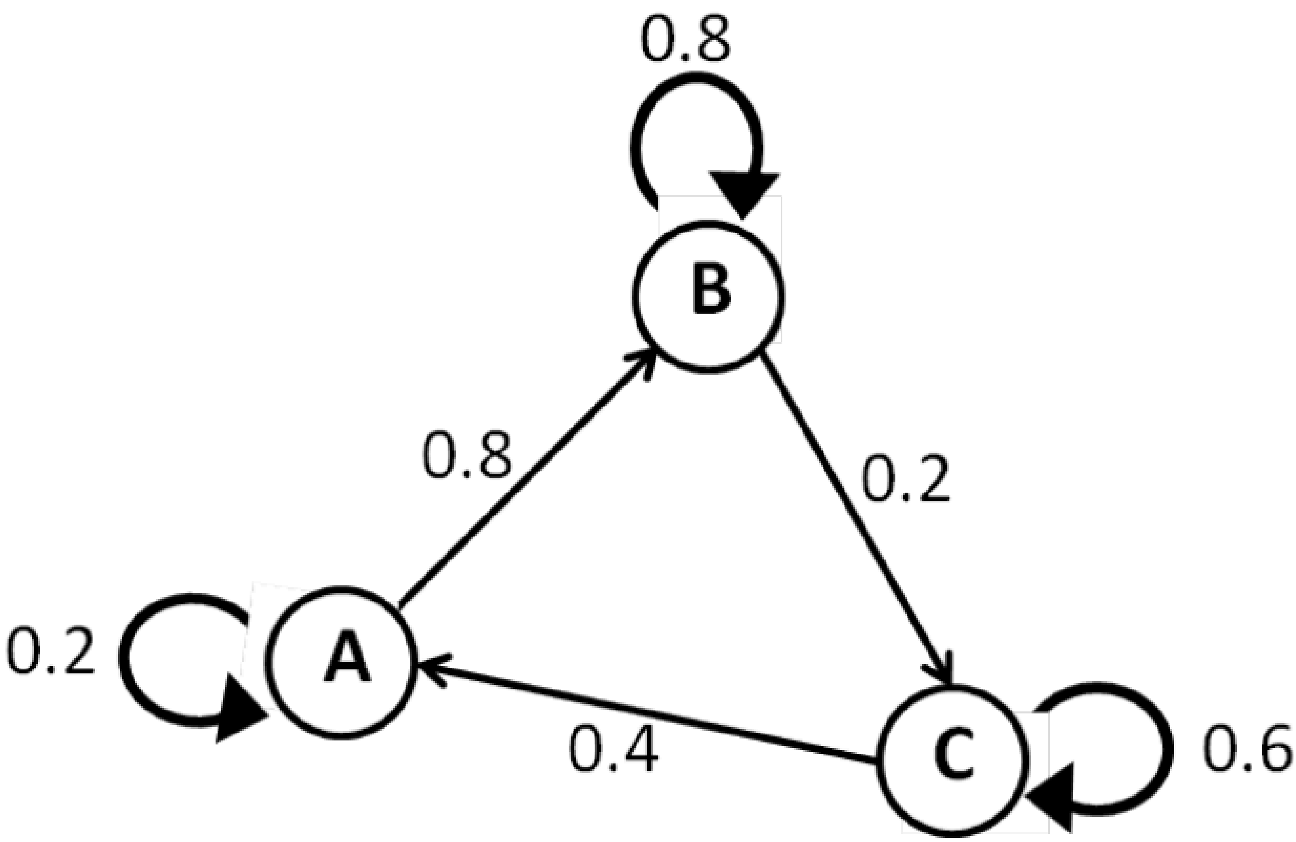
الف) نمودار حالت[1] این HMM را رسم کنید و احتمالات transition را نشان دهید.
ب) تمام مسیرهای حالت با احتمال غیر صفر را برای دنباله $O = 0, 1, 2, 0$ ارائه دهید.
ج) $P(O)$ را به کمک brute force و Forward algorithm محاسبه کنید.
د) محتمل‌ترین مسیر (Q*) چیست؟ احتمال $P(O, Q^*)$ (احتمال حرکت از این مسیر و نشر O) چیست؟ (می‌توانید از الگوریتم Viterbi استفاده کنید)
ه) برای یک دنباله مشخص O، ممکن است احتمال $P(O)$ (احتمال اینکه مدل O را در تمامی راه‌های ممکن منتشر کند) به کمک $P(O, Q^*)$ (احتمال نشر O از طریق محتمل‌ترین مسیر) تخمین زده شود. برای HMM

[1] State Diagram

مشخص شده در این سوال، آیا مقدار $P(O, Q^*)$ تخمین خوبی برای $P(O)$ می‌باشد؟ پاسخ خود را شرح دهید.

پاسخ)

الف)



ب)

AABC
ABCC
ABBC
ABCA

ج)

راه‌حل brute force:

برای بدست آوردن $P(O)$ کافیست احتمال هر یک از سری حالات بخش قبل را محاسبه کنیم. دقت کنید که در هر حالت[2] باید احتمال emission مورد نظر را هم درنظر بگیریم.

$$P(O) = P(AABC) + P(ABCC) + P(ABBC) + P(ABCA)$$
$$P(AABC) = 1.0 * 0.8 * 0.2 * 0.2 * 0.8 * 0.4 * 0.2 * 0.2 = 0.0004$$
$$P(ABBC) = 1.0 * 0.8 * 0.2 * 0.8 * 0.6 * 0.8 * 0.2 * 0.2 = 0.0025$$
$$P(ABCC) = 1.0 * 0.8 * 0.8 * 0.6 * 0.2 * 0.8 * 0.6 * 0.2 = 0.0074$$

[2] state

$$P(ABCA) = 1.0 * 0.8 * 0.6 * 0.2 * 0.8 * 0.4 * 0.8 = 0.0197$$
$$\Rightarrow P(O) = 0.0004 + 0.0025 + 0.0074 + 0.0197 = 0.0323$$

برای حل مسئله به کمک Forward Algorithm، باید به صورت بازگشتی از حالت نهایی با emission مورد نظر به حالت اول محاسبات را انجام دهیم.

$$P(O) = \sum_{s_4 \in \{a,b,c\}} P(s_4, O)$$

$$P(s_4, O) = \sum_{s_3 \in \{a,b,c\}} P(s_3, e = \{x, y, z\}).P(s_4|s_3).P(e_4 = x|s_4)$$

که در عبارت بالا، e بعنوان دنباله emissionها، s بعنوان حالت مشاهده شده، و O دنباله observationها می‌باشد.
در ادامه برای حالات ۳، ۲ و ۱ هم احتمالات مشابه را می‌نویسیم.

$$P(s_3, \{x, y, z\}) = \sum_{s_2 \in \{a,b,c\}} P(s_2, e = \{x, y\}).P(s_3|s_2).P(e_3 = z|s_3)$$

$$P(s_2, \{x, y\}) = \sum_{s_1 \in \{a,b,c\}} P(s_1, e = \{x\}).P(s_2|s_1).P(e_2 = y|s_2)$$

$$P(s_1 = A, \{x\}) = P(A)P(e_1 = x|s_1) = 1 * 0.8$$
$$P(s_1 = B, \{x\}) = 0$$
$$P(s_1 = C, \{x\}) = 0$$

حال به صورت بازگشتی مابقی احتمالات را حساب میکنیم.

$$P(s_2 = A, \{x, y\}) = 1 * 0.8 * P(s_2|s_1) * P(e_2 = y|s_2) = 1 * 0.8 * 0.2 * 0.2 = 0.032$$
$$P(s_2 = B, \{x, y\}) = 1 * 0.8 * P(s_2|s_1) * P(e_2 = y|s_2) = 1 * 0.8 * 0.8 * 0.6 = 0.384$$
$$P(s_2 = C, \{x, y\}) = 1 * 0.8 * P(s_2|s_1) * P(e_2 = y|s_2) = 0$$

$$P(s_3 = A, \{x, y, z\}) = 0.032 * 0.2 * 0.0 + 0.384 * 0.0 * 0.4 = 0$$
$$P(s_3 = B, \{x, y, z\}) = 0.032 * 0.8 * 0.4 + 0.384 * 0.8 * 0.4 = 0.13312$$
$$P(s_3 = C, \{x, y, z\}) = 0.032 * 0.0 * 0.8 + 0.384 * 0.2 * 0.8 = 0.06144$$
$$Note\ that\ O = \{x, y, z, x\}\ i.e.\ the\ observed\ sequence$$
$$P(s_4 = A, O) = 0 + 0.06144 * 0.4 * 0.8 = 0.0196608$$
$$P(s_4 = B, O) = 0.13312 * 0.8 * 0.0 + 0.06144 * 0.0 * 0.0 = 0$$
$$P(s_4 = C, O) = 0.13312 * 0.2 * 0.2 + 0.06144 * 0.6 * 0.2 = 0.0126976$$
$$\Rightarrow P(O) = 0.0196608 + 0 + 0.0126976 \simeq 0.0323$$

د)

$$P(ABCA) = 0.8 * 0.8 * 0.6 * 0.2 * 0.8 * 0.4 * 0.8 = 0.0197$$

ه)

در این حالت احتمال محتمل‌ترین مسیر تقریبا دو سوم احتمال کل خواهد بود که تقریب قابل قبولی است. در برخی حالات این اتفاق نمی‌افتد و مقدار تقریبی بسیار کمتر از مقدار واقعی می‌باشد.

Naive Bayes

طراح: اولدوز نیساری

به سوالات زیر در خصوص مفهوم naive bayes پاسخ تشریحی دهید.

1) تفاوت بین diagnostic probability و causal probability را در naive bayes بیان کنید و با یک مثال توضیح دهید که چگونه naive bayes این دو را به یکدیگر مرتبط می‌کند؟
2) در یک مطالعه پزشکی ، متوجه شده‌اند که یک نوع بیماری تنها یک درصد از جامعه را تحت تاثیر قرار می‌دهد. یک تست تشخیصی انجام شده است. اگر فردی مبتلا باشد، این آزمایش به احتمال 99 درصد او را بیمار تشخیص می‌دهد ، اما اگر فردی به بیماری مبتلا نباشد ، به احتمال 5 درصد او را به اشتباه بیمار تشخیص می‌دهد . با توجه به داده‌های این مسئله توضیح دهید احتمال مثبت بودن نتیجه آزمایش در صورت ابتلا چه نقشی در محاسبه احتمال ابتلا در صورت مثبت بودن دارد؟ آیا وقتی از naive bayes برای استنتاج درباره این مسئله استفاده می‌کنیم، فرضی درباره استقلال پیشامدها داریم؟ احتمال بیمار نبودن را به شرط منفی بودن نتیجه تست را محاسبه کنید.

پاسخ)

1) در naive bayes، احتمالی که مبنای استنتاج ما است، diagnostic probability است و احتمالی که آن را استنتاج می‌کنیم causal probability است. naive bayes رابطه‌ای به ما ارائه می‌دهد که بر اساس آن می‌توانیم استنتاج را انجام دهیم . فرض کنید ما احتمال دندان درد به شرط پوسیدگی را داشته باشیم و بخواهیم احتمال پوسیدگی به شرط دندان درد را محاسبه کنیم. در این جا احتمال دندان درد به شرط پوسیدگی diagnostic probability است و احتمال پوسیدگی به شرط دندان درد causal probability است که به کمک رابطه زیر محاسبه می‌کنیم:

$$p(cavity \mid toothache) = \frac{P(toothache \mid cavity)\, P(cavity)}{P(toothache)}$$

2) اگر بخواهیم احتمال مثبت بودن نتیجه آزمایش در صورت ابتلا را محاسبه کنیم، طبق naive bayes این احتمال می‌شود:

اگر مثبت بودن نتیجه آزمایش به شرط ابتلا را causal probability در نظر بگیریم، می‌توانیم آن را بر حسب diagnostic probability اش یعنی احتمالش محاسبه کنیم در این صورت طبق فرمول:

$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$ نیاز داریم که احتمال مثبت بودن آزمایش تشخیصی را محاسبه کنیم و آن را به صورت تاثیر مستقیم در محاسبه‌مان تاثیر دهیم.

بله یک فرض مهم داریم که ابتلا یک فرد به بیماری و تشخیص بیماری‌اش توسط آزمایش از سایر feature ها و عوامل مستقل هستند.

$$P(not\ sick \mid negative) = \frac{P(negative \mid not\ sick)\, P(not\ sick)}{P(negative)}$$

$$P(negative \mid not\ sick) = 1 - p(positive \mid not\ sick) = 1 - 0.05 = 0.95$$

$P(not\ sick) \ = \ 0.99$

$P(negative\ ) \ = \ (1 \ - \ P(positive \mid sick)). P(sick) \ + \ P(negative \mid not\ sick\ )\ P(not\ sick) \ =$
$(1 \ - \ 0.99). (0.01) \ + \ 0.95. 0.99$

$P(not\ sick \mid negative\ ) \ = \ \frac{0.95\ .\ 0.99}{0.01*0.01\ +\ 0.95.0.99}$

Decision Trees

طراح: صادق فاضلی

آقای وحیدی که صاحب یک فروشگاه اینترنتی است، چندین کد تخفیف را از طریق پیامک برای تعدادی از مشتری‌های سابق فروشگاه ارسال کرده بود. بعضی از مشتریانی که کد تخفیف دریافت کرده بودند، از کد خود استفاده کرده و با خرید خود، آقای وحیدی را خوشحال کرده بودند. داده‌های مربوط مشتریان و استفاده آنها از کد تخفیف در جدول "داده‌های آموزش" قابل مشاهده است.

از آنجا که ارسال پیامک هزینه دارد، آقای وحیدی می‌خواهد بهینه عمل کرده و کدهای تخفیف جدید را فقط برای مشتریانی ارسال کند که انتظار می‌رود از فروشگاه خرید کنند. به همین دلیل آقای وحیدی از شما می‌خواهد یک Classifier طراحی کنید که بر اساس اطلاعات مشتری، استفاده کردن او از کد تخفیف را پیش‌بینی کند.

داده‌های آموزش					
	جنسیت	متاهل	سن	سابقه خرید در ماه گذشته	استفاده از تخفیف
۱	زن	بله	<۲۵	بله	✓
۲	مرد	خیر	۲۵-۴۰	بله	✓
۳	مرد	بله	۴۰<	خیر	×
۴	زن	خیر	<۲۵	بله	✓
۵	مرد	بله	۲۵-۴۰	خیر	×
۶	زن	بله	۲۵-۴۰	خیر	✓
۷	مرد	بله	۲۵-۴۰	بله	×
۸	مرد	بله	۲۵-۴۰	خیر	×
۹	زن	خیر	<۲۵	خیر	✓
۱۰	مرد	بله	۴۰<	خیر	✓
۱۱	زن	خیر	۲۵-۴۰	بله	✓

دادههای آموزش					
	جنسیت	متاهل	سن	سابقه خرید در ماه گذشته	استفاده از تخفیف
۱۲	مرد	خیر	۴۰<	بله	×
۱۳	زن	خیر	<۲۵	خیر	✓
۱۴	مرد	بله	<۲۵	بله	×
۱۵	زن	خیر	۴۰<	خیر	✓

قسمت اول

یک Classifier بر اساس Information Gain و با عمق ۳ (با احتساب ریشه و برگ‌ها) برای پیش‌بینی استفاده از کد تخفیف بسازید. از جدول "داده‌های آموزش" استفاده کنید. درخت نهایی و مراحل محاسبات خود را بنویسید.

قسمت دوم

با استفاده از Classifier ساخته شده در قسمت اول، "داده‌های آزمون" را پیش‌بینی کنید. سپس ماتریس درهم‌ریختگی (Confusion Matrix) را رسم کرده و Accuracy و Precision و Recall را محاسبه کنید.

داده‌های آزمون					
	جنسیت	متاهل	سن	سابقه خرید در ماه گذشته	استفاده از تخفیف
۱	مرد	بله	۴۰<	بله	✓
۲	زن	خیر	<۲۵	خیر	×
۳	مرد	بله	۲۵-۴۰	خیر	×
۴	زن	خیر	<۲۵	بله	✓
۵	مرد	خیر	۴۰<	بله	✓
۶	مرد	بله	۲۵-۴۰	خیر	×
۷	زن	خیر	<۲۵	خیر	✓
۸	زن	بله	<۲۵	بله	✓

قسمت سوم

به سوالات زیر به صورت تشریحی پاسخ دهید:

1) Classifier های درخت تصمیم چه زمانی دچار بیش‌برازش (Overfitting) می‌شوند؟ دلیل آن چیست؟
2) دو روش برای جلوگیری از بیش‌برازش (Overfitting) در Classifier های درخت تصمیم ارائه دهید.

پاسخ)

قسمت اول

ابتدا آنتروپی اولیه را محاسبه می‌کنیم.

$$E(Using\ Discount) = -\frac{9}{15}log(\frac{9}{15}) - \frac{6}{15}log(\frac{6}{15}) = 0.971$$

سپس آنتروپی و میانگین وزن‌دار ویژگی‌ها را محاسبه می‌کنیم.

جنسیت:

$$E(man) = -\frac{2}{8}log(\frac{2}{8}) - \frac{6}{8}log(\frac{6}{8}) = 0.81125$$

$$E(woman) = -\frac{7}{7}log(\frac{7}{7}) - 0 = 0$$

$$AE(Gender) = \frac{8}{15}E(man) + \frac{7}{15}E(woman) = 0.43266$$

تاهل:

$$E(married) = -\frac{3}{8}log(\frac{3}{8}) - \frac{5}{8}log(\frac{5}{8}) = 0.95443$$

$$E(single) = -\frac{6}{7}log(\frac{6}{7}) - \frac{1}{7}log(\frac{1}{7}) = 0.59168$$

$$AE(Marital\ Status) = \frac{8}{15}E(married) + \frac{7}{15}E(single) = 0.78514$$

سن:

$$E(< 25) = -\frac{4}{5}log(\frac{4}{5}) - \frac{1}{5}log(\frac{1}{5}) = 0.4644$$

$$E(25 - 40) = -\frac{3}{6}log(\frac{3}{6}) - \frac{3}{6}log(\frac{3}{6}) = 1$$

$$E(40 >) = -\frac{2}{4}log(\frac{2}{4}) - \frac{2}{4}log(\frac{2}{4}) = 1$$

$$AE(Age) = \frac{5}{15}E(< 25) + \frac{6}{15}E(25 - 40) + \frac{6}{15}E(40 >) = 0.82146$$

سابقه خرید:

$$E(yes) = -\frac{4}{7}log(\frac{4}{7}) - \frac{3}{7}log(\frac{3}{7}) = 0.98525$$

$$E(no) = -\frac{5}{8}log(\frac{5}{8}) - \frac{3}{8}log(\frac{3}{8}) = 0.95443$$

$$AE(Purchase\ History) = \frac{7}{15}E(yes) + \frac{8}{15}E(no) = 0.96881$$

تفاضل آنتروپی اولیه با میانگین وزن‌دار هر ویژگی، برابر با information gain آن ویژگی است. در نتیجه جنسیت بیشترین information gain را دارد پس به عنوان اولین ویژگی برای تقسیم درخت انتخاب می‌شود.

در مرحله بعد برای دو حالت زن و مرد، مراحل بخش قبل را تکرار می‌کنیم.

مرد

$$E(man) = -\frac{2}{8}log(\frac{2}{8}) - \frac{6}{8}log(\frac{6}{8}) = 0.81125$$

تاهل:

$$E(married) = -\frac{1}{6}log(\frac{1}{6}) - \frac{5}{6}log(\frac{5}{6}) = 0.650025$$

$$E(single) = -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2}) = 1$$

$$AE(Marital\ Status) = \frac{6}{8}E(married) + \frac{2}{8}E(single) = 0.73751875$$

سن:

$$E(< 25) = -0 - \frac{1}{1}log(\frac{1}{1}) = 0$$

$$E(25 - 40) = -\frac{1}{4}log(\frac{1}{4}) - \frac{3}{4}log(\frac{3}{4}) = 0.81125$$

$$E(40 >) = -\frac{1}{3}log(\frac{1}{3}) - \frac{2}{3}log(\frac{2}{3}) = 0.918333$$

$$AE(Age) = \frac{1}{8}E(< 25) + \frac{4}{8}E(25 - 40) + \frac{3}{8}E(40 >) = 0.7499$$

سابقه خرید:

$$E(yes) = -\frac{1}{4}log(\frac{1}{4}) - \frac{3}{4}log(\frac{3}{4}) = 0.81125$$

$$E(no) = -\frac{1}{4}log(\frac{1}{4}) - \frac{3}{4}log(\frac{3}{4}) = 0.81125$$

$$AE(Purchase\ History) = \frac{4}{8}E(yes) + \frac{4}{8}E(no) = 0.81125$$

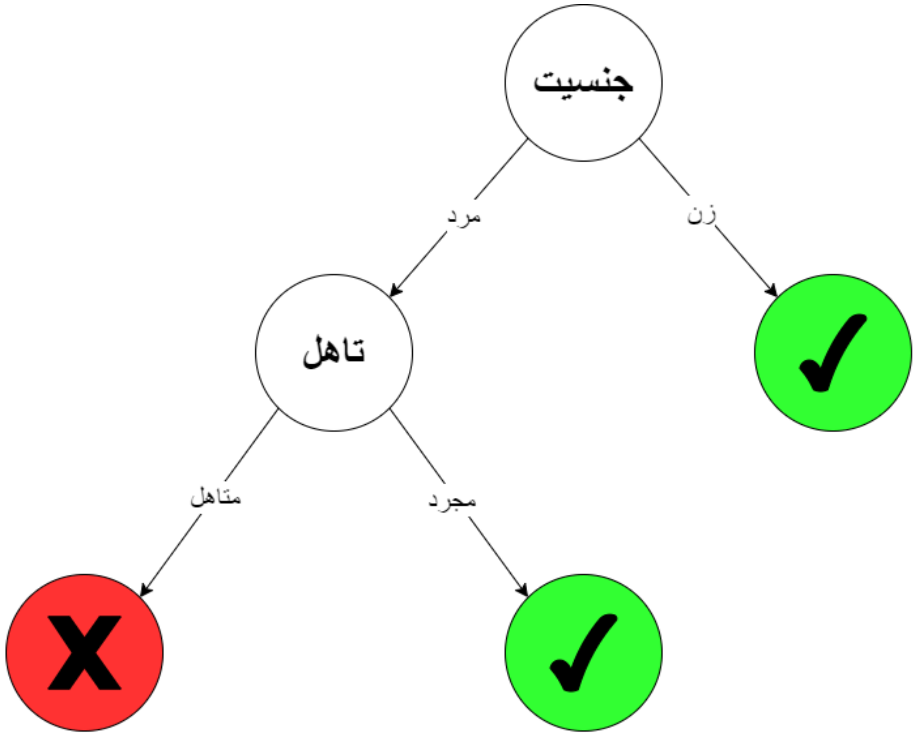
ویژگی تاهل بیشترین information gain را دارد پس برای ادامه تقسیم درخت استفاده می‌شود. در صورت داشتن تاهل، تصمیم‌گیری به این صورت خواهد بود که فرد از تخفیف استفاده نمی‌کند زیرا داده‌ها به نسبت 5 به 1 به سمت استفاده نکردن از تخفیف هستند، اما در صورت مجرد بودن داده‌ها به نسبت مساوی تقسیم شده‌اند و تصمیم با ماست که چگونه دسته‌بندی کنیم. از آنجایی که در این بخش داشتن تاهل را معادل استفاده نکردن از تخفیف دسته‌بندی کردیم، مجرد بودن را معادل استفاده از تخفیف در نظر می‌گیریم.

زن

$$E(woman) = -\frac{7}{7}log(\frac{7}{7}) - 0 = 0$$

آنتروپی این ویژگی 0 است پس نیازی به محاسبه مجدد آنتروپی فرزندان نیست و این شاخه داده‌ها را در دسته مربوط استفاده از تخفیف = ✓ طبقه‌بندی می‌کند.

طبق نتایج بدست آمده، درخت تصمیم به صورت زیر رسم می‌شود:



قسمت دوم

		Actual	
		✓	×
Predicted	✓	4	1
	×	1	2

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{6}{8} = 75\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{4}{5} = 80\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{4}{5} = 80\%$$

قسمت سوم

1) طبقه‌بندهای درخت تصمیم مستعد overfitting هستند. زمانی که به درخت تصمیم اجازه دهیم که تا بیشترین عمق ممکن رشد کند، درخت بر روی داده‌های آموزش overfit می‌شود. overfitting در درخت تصمیم به این دلیل رخ می‌دهد که درخت بیش از حد پیچیده شده و بجای دریافت الگوی موجود در داده‌ها، به دریافت نویز داده آموزش می‌پردازد. این اتفاق باعث می‌شود که عملکرد generalization مدل افت کند و مدل در برابر داده‌های دیده نشده ضعیف عمل کند.

2) روش‌های متفاوتی برای جلوگیری از overfitting در درخت‌های تصمیم وجود دارد که در اینجا به صورت مختصر به سه مورد اشاره می‌کنیم:

1- Pruning: در این روش با حذف کردن بخش‌هایی از درخت تصمیم از رشد بیش از حد آن جلوگیری می‌کنیم. این روش باعث می‌شود که بخش‌ها و شاخه‌هایی از درخت تصمیم که قابلیت طبقه‌بندی داده‌ها را ندارند از درخت حذف شوند. این کار به دو حالت می‌تواند انجام شود: Pre-Pruning و Post-Pruning . روش Pre-Pruning در زمان ایجاد درخت، از شکل‌گیری شاخه‌هایی با ارزش داده‌ای کم جلوگیری می‌کند و نمی‌گذارد درخت به عمق حداکثری خود برسد. در مقابل، Post-Pruning اجازه می‌دهد که درخت تا عمق نهایی رشد کند و سپس به حذف کردن شاخه‌ها و گره‌های overfit شده می‌پردازد.

2- Limiting Maximum Depth: در این روش با استفاده از hyperparameter های درخت، برای آن محدودیت حداکثر عمق تعیین می‌کنیم که جلوی رشد بیش از اندازه درخت را بگیرد. این کار باعث می‌شود که صرفا ویژگی‌های دارای gain اطلاعاتی بالا در ایجاد درخت شرکت کنند.

3- Feature Selection: در این روش، ابتدا ویژگی‌هایی را که وابستگی داده‌ای بیشتری با ویژگی هدف دارند را جدا کرده و سپس فقط با استفاده از آنها درخت را ایجاد می‌کنیم. این کار باعث می‌شود که پیچیدگی مدل کم شود و از overfitting جلوگیری شود.

Convolutional Neural Networks

طراح: علی محمدی

قسمت اول

شما در حال آموزش دادن یک Convolutional Neural Networks بر روی دیتاست ImageNet هستید، و شما در حال فکر کردن به این موضوع هستید که برای optimization function خود از gradient descent استفاده کنید. کدام گزینه/گزینه‌های زیر درست می‌باشد. توضیح کوتاه کافی است.

1. ممکن است که Stochastic Gradient Descent سریع‌تر از Batch Gradient Descent همگرا شود.
2. ممکن است که Mini Batch Gradient Descent سریع‌تر از Stochastic Gradient Descent همگرا شود.
3. ممکن است که Mini Batch Gradient Descent سریع‌تر از Batch Gradient Descent همگرا شود.
4. ممکن است که Batch Gradient Descent سریع‌تر از Stochastic Gradient Descent همگرا شود.

قسمت دوم

شما با یک تابع غیر خطی مواجه می‌شوید که اگر ورودی آن غیرمنفی باشد، عدد 1 را پاس می کند و درغیر اینصورت به 0 را می‌دهد، یعنی:

$$f(x) = \begin{cases} 1 & | \; x >= 0 \\ 0 & | \; x < 0 \end{cases}$$

یکی از دوستان توصیه می کند که از این غیر خطی بودن در شبکه عصبی کانولوشنال خود با بهینه ساز Adam استفاده کنید. آیا به توصیه آنها عمل می کنید؟ چرا و چرا نه؟

پاسخ)

قسمت اول

۱ و ۲ و ۳

قسمت دوم

خیر، چرا که اگر چه از نظر فنی یک تابع غیر خطی است(به طور ویژه یک تابع غیر خطی ناپیوسته تابع پله)، که یعنی گرادیان آن در همه نقاط به جز مبدا آن صفر است. پس تقریبا هیچ گرادیانی در حین back prop برنمی‌گردد، در حالی که که هنگام بهینه سازی با ADAM یا هر بهینه ساز دیگر، descent-based بودن بسیار مهم است.